心理学报 2019, Vol. 51, No. 6, 734−746
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2019.00734

一种基于多阶认知诊断模型测评科学素养的方法*

詹沛达 于照辉 李菲茗 王立君
(浙江师范大学教师教育学院，金华 321004)

摘 要 科学素养是指作为一名有反思意识的公民所具有的解决科学问题和运用科学理念的能力。为实现在认知诊断中对科学素养的测评，本文基于 PISA 2015 科学素养测评框架首次提出科学素养包含的三阶潜在结构，使用新提出的多阶认知诊断模型对 PISA 2015 科学测评数据进行分析，并通过模拟研究探究新模型的心理测量学性能。结果表明：(1)新模型能够较好地分析包含三阶潜在结构的科学素养；(2)科学知识对科学素养的影响最大，科学背景次之，科学能力的影响最小；(3)全贝叶斯 MCMC 算法能够为新模型提供较精准的参数估计。

关键词 科学素养；认知诊断；PISA；DINA 模型
分类号 B841

1 引言

“科学技术推动了生产力的发展、经济的繁荣和社会的进步，促进了人们的生产方式、生活方式和思维方式的变革。科学技术的快速发展对每一位公民的科学素养提出了新的要求” (中华人民共和国教育部，2017)。实际上，关于如何提高个体或公民的科学素养是一个交叉学科问题，它一直以来都是科学教育、教育心理学和学习科学等学科领域的学者们共同关注的重难点。科学素养是一个不断发展的概念，它的内涵和界定方式会随时代发展而发生改变(see Miller, 1983; OECD, 2006)。2017 年，《义务教育小学科学课程标准》将“科学素养”定义为“了解必要的科学技术知识及其对社会与个人的影响，知道基本的科学方法，认知科学本质，树立科学思想，崇尚科学精神，并具备一定的运用它们处理实际问题、参与公共事务的能力”，从本质上讲，该定义就是说“科学素养是指作为一名有反思意识的公民所具有的解决科学问题和运用科学理念的能力” (OECD, 2016)。

为实现对科学素养的客观测评，国际学生评估项目(Programme for International Student Assessment, PISA)在 2015 年把科学素养的内涵划分为科学能力(Competencies)、科学知识(Knowledge)、科学背景(Contexts)和科学态度(Attitudes)四个相互关联的维度，并给出了相应的测评或评估框架，见图 1。这就要求学生在一定的科学背景中，根据自己的科学态度，运用科学知识来解决科学问题，从而展现出自己的科学能力(刘克文，李川，2015)。PISA 2015 测评框架是在 PISA 2006 科学测评框架(OECD, 2006)的基础上修订而来的，其发展主要体现在对科学知识维度的更详细划分。科学测评框架的逐步完善，是在实践基础上不断重新认识科学素养的结果。可以说，PISA 2015 科学素养测评框架是目前最新最具可操作性的科学素养测评框架。

除具有可操作性的测评框架外，一个适宜测评方法也同样重要。适宜的测评方法应能够匹配测评框架，并能够实现对科学素养客观且准确的评价。

收稿日期: 2018-09-21
* 国家自然科学基金青年基金项目(31600908)、浙江省自然科学基金项目(LY16C090001)、教育部人文社会科学研究青年基金项目(19YJC190025)和浙江省教育科学规划重点课题(2019SB112)资助。
通信作者：詹沛达, E-mail: pdzhan@gmail.com

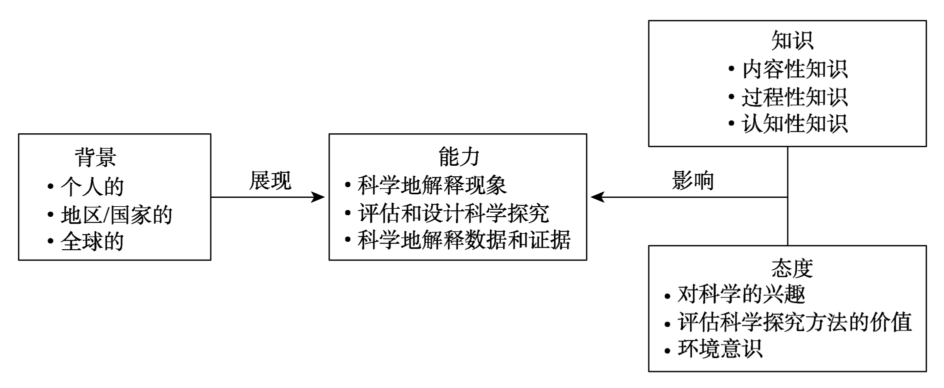


图 1　PISA 2015 科学素养测评框架(来源：OECD (2016)第 23 页图 2.2).

然而，目前国内外已有研究绝大多数只是对公民或中小学学生科学素养的问卷调查(e.g., Roos, 2014; 高宏斌, 2011; 秦浩正, 钱源伟, 2008)，这仅是对科学素养整体现状的大致了解。而且这些调查多采用自我报告法，主观性较强，存在一定的社会赞许性。仅有少许研究关注到了对科学素养的测评(e.g., 胡咏梅, 杨素红, 卢珂, 2012)。除研究方法有待改进外，目前绝大多数研究所使用的测评/数据分析方法和理论也较为落后，仍以经典测量理论为主(e.g., Roos, 2014; 任磊, 张超, 何薇, 2013)，仅有个别研究使用到了题目作答理论(item response theory, IRT)模型(e.g., 胡咏梅等, 2012)。另外，需要强调的是尽管 PISA 为科学素养建构了多维结构，但数据分析时仍使用了单维 IRT 模型(OECD, 2017)。即 PISA 现有的测评方法并不匹配测评框架，其主要原因之一是因为PISA 更关注的是国家/经济体的整体现状而非个体参与者，所以对个体使用一个笼统的单维潜在特质可以简化整体研究的复杂性。而当把个体视为测评主体时，就需要更复杂的测评方法(e.g., Zhan, Jiao, & Liao, 2018)。综上所述，为在 PISA 2015 科学素养测评框架下实现对科学素养客观且准确的测评，需要尝试从新的视角切入，使用或开发更适宜的测评方法。

近些年，随着认知心理学的发展，研究者们逐渐发现被试在完成某项任务时常需要多种能力的相互配合，因此，早期心理测量模型中的单维性假设并不符合实际(Reckase, 2009; Wang & Chen, 2004; 康春花, 辛涛, 2010; 詹沛达, 王文中, 王立君, 2013)。另外，除了简单的总分外，人们也希望能从被试的实际作答情况中获得更丰富的信息，以便对被试做出更客观的评价和补救。基于此，认知诊断测评(cognitive diagnostic assessment, CDA)在近一二十年内受到了国内外学者的更多关注(Rupp, Templin, & Henson, 2010; 涂冬波, 蔡艳, 丁树良, 2012)。CDA 是指在心理与教育测量学中对个体认知过程、加工技能或知识结构(统称为属性)的诊断性测评。作为一种将形成性评价和终结性评价相结合的综合评价形式(詹沛达, 陈平, 边玉芳, 2016)，CDA 的初衷是通过测评个体对属性的掌握状态为教师或干预者提供诊断反馈报告，进而帮助他们实施补救教学或有针对性的干预(Zhan et al., 2018)。CDA 改变了以往评价方法重结果、轻过程的弊端，符合当前我国一些教育政策导向。比如:《基础教育课程改革纲要(试行)》中“改变课程评价过分强调甄别与选拔的功能，发挥评价促进学生发展、教师提高和改进教学实践的功能”的具体目标。因此，如何在 CDA 中实现对科学素养的测评是一个兼具理论意义和实践意义的议题。

下文中，我们首先将对 PISA 2015 科学素养测评框架做进一步解读，明确该框架所包含的三阶潜在结构；其次，对现有的高阶认知诊断模型(higher-order cognitive diagnosis model; HO-CDM)进行介绍并阐明其局限性；然后，提出一种新的多阶认知诊断模型(multi-order CDM; MO-CDM)，以期在 CDA 中满足对三阶或更高阶潜在特质的分析需求，并匹配 PISA 2015 科学素养测评框架，实现对科学素养的准确测评。再然后，我们以 PISA 2015 科学测评数据分析为例来说明新模型的现实可应用性，并对数据分析结果进行解读。最后，通过一个模拟研究来探究新模型的参数估计返真性。

2　科学素养包含的三阶潜在结构

PISA 2015 认为科学素养的核心是科学能力，而科学能力的展现需要在特定的科学背景下辅以足够的科学知识，并受到科学态度的影响。这 4 个维度相辅相成，共同组成了科学素养，即科学素养

是科学能力、科学知识、科学背景和科学态度的高阶/高位概念，个体科学素养的高低决定了他在这 4 个维度方面的表现情况。进一步，根据《PISA 2015 测评与分析框架》(OECD, 2016)：

(1) 科学能力又被细分为 3 种子能力，分别是科学地解释现象、评估和设计科学探究和科学地解释数据和证据。即科学能力是 3 这子能力的高阶概念，个体科学能力的高低决定了其 3 项子能力的高低；

(2) 科学知识又被细分为 3 种子知识，分别是内容性知识、程序性知识和认知性知识。即科学知识是这 3 种子知识的高阶概念，个体对科学知识的掌握程度决定了其对 3 种子知识的掌握程度；

(3) 科学背景又被细分为 3 种子背景，分别是个人的、当地/国家的和全球的。即科学背景是这 3 个子背景的高阶概念，个体对科学背景的熟悉程度影响着其对 3 种子背景的熟悉程度；

(4) 科学态度又被细分为 3 种子态度，分别是对科学的兴趣、评估科学探究方法的价值和环境意识。即科学态度是这 3 种子态度的高阶概念，个体的科学态度影响其 3 种子态度。

综上所述，基于 PISA 2015 科学素养测评框架，科学素养包含三阶潜在结构，如图 2 所示。其中，第三阶潜在特质为科学素养，是 PISA 2015 科学素养测评框架中的最高阶概念；第二阶潜在特质包括：科学能力、科学知识、科学背景和科学态度，是该测评框架中的 4 个主要概念；而第一阶潜在特质为科学地解释现象、评估和设计科学探究等 12 项，是该测评框架中的低阶概念。

为在 CDA 中实现对科学素养的测评，需要一种能够分析科学素养三阶潜在结构的 CDM。鉴于目前尚未有 CDM 能够处理三阶潜在结构，这就需要我们建构新的模型，以期满足测评需求。

3 多阶认知诊断模型

3.1 高阶认知诊断模型及其局限性

在心理学和教育学中，潜在特质除了可能存在多维性外，还可能进一步存在层阶关系，这被称为高阶(层阶)潜在特质，比如，图 2 所示的科学素养所包含的三阶潜在结构；再比如，韦氏成人智力量表中也测量了三阶潜在特质：第一阶中包含了 13 个子测验并分别测量了一种潜在特质，在第二阶中这 13 种潜质就被归为 4 种外延更广的潜在特质(言语能力、知觉推理、工作记忆和信息加工速度)，而在第三阶中这 4 种潜在特质又包含在一般智力之中(Ryan & Schnakenberg-Ott, 2003)。

高阶潜在特质的概念是建构在多维潜在特质概念之上的，用于描述多个潜在特质之间可能存在的结构关系。基于此，研究者们开发了两类不同的高阶心理测量模型(陈飞鹏，詹沛达，王立君，陈春晓，蔡毛, 2015)：基于多维 IRT 模型建构的高阶 IRT 模型(de la Torre & Song, 2009; Huang, Wang, Chen, & Su, 2013; Rijmen, Jeon, von Davier, & Rabe-Hesketh, 2014)和基于 CDM 建构的高阶认知诊断模型(HO-CDM) (de la Torre & Douglas, 2004; Templin, Henson, Templin, & Roussos, 2008; Zhan, Wang, & Li, in press)，本文聚焦于后者。

在 CDA 中，鉴于被试对属性的掌握可能受到一个(或多个)更高阶的潜在特质的影响且为减少参数估计的数量，de la Torre 和 Douglas(2004)提出了高阶潜在结构模型

$$\operatorname{logit}(P(\alpha_k = 1 \mid \theta_n)) = \lambda_{1k}\theta_n - \lambda_{0k}\,. \qquad (1)$$

式中，$\operatorname{logit}(x) = \log\left(\frac{x}{1-x}\right)$；$P(\alpha_{nk} = 1 \mid \theta_n)$为给定第

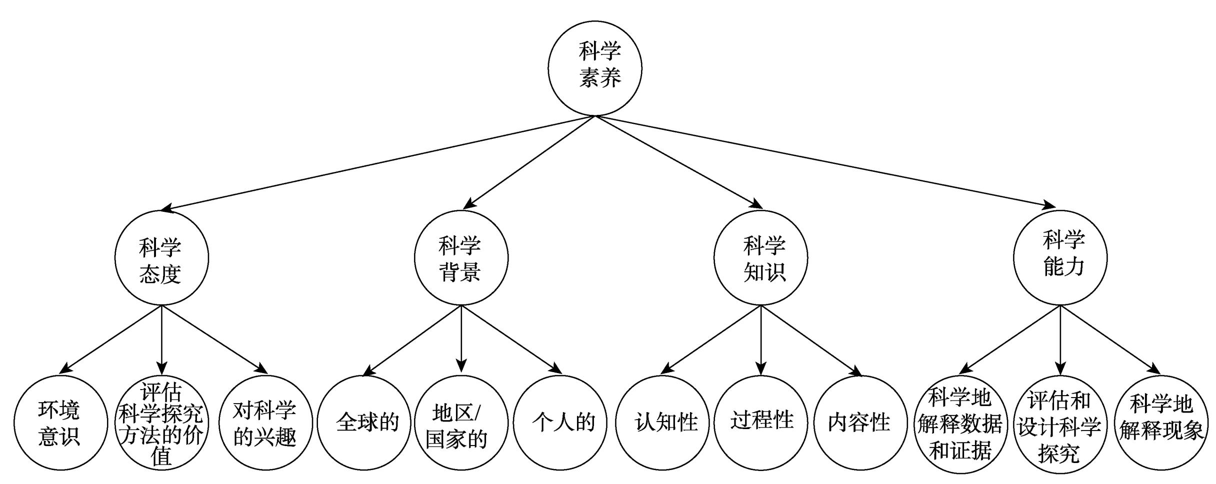


图 2 PISA 2015 科学素养所包含的三阶潜在结构

二阶潜在特质 θ_n 后被试 n 掌握属性 k 的概率；λ_{0k} 为属性 k 的难度参数，λ_{1k} 为属性 k 的区分度参数。式(1)所描述的潜在结构见图 3。式(1)是潜在结构模型，将它们与测量模型相结合即可得到 HO-CDM。比如，将它们与 DINA 模型(Junker & Sijtsma, 2001; Macready & Dayton, 1977)相结合即可得到高阶 DINA (HO-DINA)模型。限于高阶潜在结构模型的理论局限，HO-DINA 模型只能处理包含二阶潜在结构的数据，无法实现对科学素养所包含的三阶潜在结构的测评，不满足本研究的需求。

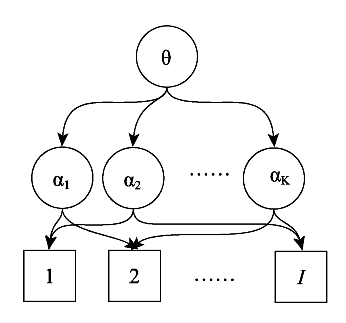


图 3　CDA 中二阶潜在特质与属性间的关系示例图
注：θ 为第二阶潜在特质；α 为(第一阶)属性；K 为总属性数量；I 为总题目数量

3.2　多阶认知诊断模型的建构

3.2.1　多阶潜在结构模型(MO-LSM)

针对目前缺乏可处理三阶或更高阶潜在结构的 CDM 这一问题，本研究借鉴高阶 IRT 模型的建模思路，把线性潜在结构模型引入到当前的二阶潜在结构模型(式(1))之上，提出多阶潜在结构模型(multi-order latent structural model; MO-LSM)。首先，假设潜在特质存在多阶结构，$\theta_{nm}^{(h)}$ 表示被试 n 在第 h ($h \geqslant 2$)阶中的第 m 个潜在特质，则 $\theta_{nm}^{(h)}$ 与更高阶的潜在特质 $\boldsymbol{\theta}_n^{(h+1)}$ 之间的线性潜在结构关系可被描述为：

$$\theta_{nm}^{(h)} = \boldsymbol{\gamma}_m'^{(h)}\boldsymbol{\theta}_n^{(h+1)} + \varepsilon_{nm}^{(h)} = \sum_{p=1}^{P}\gamma_{mp}^{(h)}\theta_{np}^{(h+1)} + \varepsilon_{nm}^{(h)}, \quad (2)$$

式中，$\boldsymbol{\gamma}_m'^{(h)}$ 为第 h 阶回归向量；$\varepsilon_{nm}^{(h)}$ 为第 h 阶中的第 m 个潜在特质的残差；$\theta_{np}^{(h+1)}$ 为被试 n 在第 $h+1$ 阶中的第 p 个潜在特质。需要说明的是，除了线性关系外，式(2)也可以修改为非线性关系(e.g., 多项式)，但鉴于心理学研究中通常假设潜变量之间为线性关系(e.g., 结构方程模型)，且为降低模型复杂性，本研究暂只关注线性关系(de la Torre & Song, 2009; Huang et al., 2013; Rijmen et al., 2014)。将式(2)引入式(1)中即可得到 MO-LSM：

$$\text{logit}(P(\alpha_{nk}=1|\boldsymbol{\theta}_n^{(h)})) = \sum\nolimits_{m=1}^{M}\lambda_{1mk}\theta_{nm}^{(h)} - \lambda_{0k} = \sum\nolimits_{m=1}^{M}\lambda_{1mk}(\sum\nolimits_{p=1}^{P}\gamma_{mp}^{(h)}\theta_{np}^{(h+1)} + \varepsilon_{nm}^{(h)}) - \lambda_{0k}. \quad (3)$$

基于条件独立性假设，MO-LSM 假设当给定更高一阶的潜在特质时，各低阶潜在特质之间相互独立。需要说明的是，尽管式(3)在理论上能够处理多阶的潜在特质，但考虑到现实测验情境中出现四阶潜在特质的可能性已经较小，且为匹配 PISA 2015 科学素养所包含的三阶潜在结构，本研究聚焦于仅包含 1 个第三阶潜在特质的三阶潜在结构模型，如图 4，该模型可被描述为：

$$\text{logit}(P(\alpha_{nk}=1|\boldsymbol{\theta}_n^{(2)})) = \sum_{m=1}^{M}\lambda_{1mk}\theta_{nm}^{(2)} - \lambda_{0k} = \sum_{m=1}^{M}\lambda_{1mk}(\gamma_m^{(2)}\theta_n^{(3)} + \varepsilon_{nm}^{(2)}) - \lambda_{0k}. \quad (4)$$

为使模型可识别，设定 $\theta_n^{(3)} \sim N(0, 1)$ 且 $\varepsilon_{nm}^{(2)} \sim N(0, 1-\gamma_m^{(2)^2})$，进而有 $\theta_{nm}^{(2)} \sim N(0, 1)$，此时，任意两个第二阶潜在特质之间的相关系数等于 $\gamma_m^{(2)} \times \gamma_{m'}^{(2)}$。当 $\gamma_m^{(2)} = 1$ 时，有 $\varepsilon_{nm}^{(2)} = 0$，则式(4)退化为式(1)。

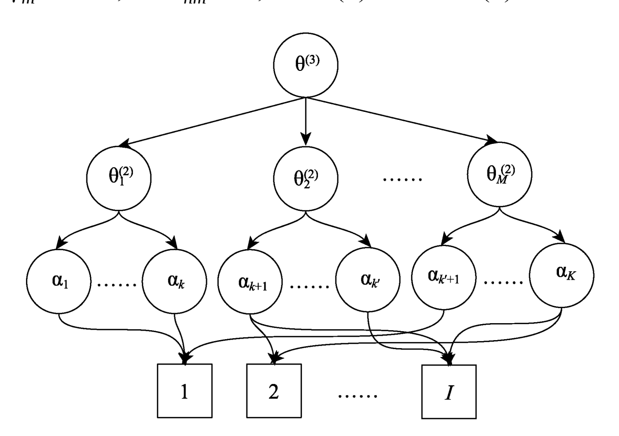


图 4　CDA 中第三阶潜在特质与属性间的关系示例图.
注：$\theta^{(3)}$为第三阶潜在特质；$\theta^{(2)}$为第二阶潜在特质；α 为(第一阶)属性；K 为总属性数量；I 为总题目数量。

3.2.2　MO-DINA 模型

通常，CDM 由两部分组成：测量模型和潜在结构模型(Rupp et al., 2010)，前者定义了被试作答题目的正确概率，后者描述了属性之间的结构关系。在 3.2.1 中，我们已经定义了 MO-LSM，为提高参数估计的精度和效率，我们选用引入题目内特征依赖性的贝叶斯 DINA 模型(Zhan, Jiao, Liao, & Bian, 2018)作为测量模型，模型详述见附录。

本研究采用全贝叶斯马尔可夫链蒙特卡洛(MCMC)算法来实现对 MO-DINA 模型的参数估计，并基于 JAGS 软件(Version 4.3.0)实现。各待估计参数的先验分布详见附录，相应的 JAGS 代码也可向作

者索取。关于如何使用 JAGS 实现对贝叶斯 CDM 的参数估计，可参阅 Zhan、Jiao、Man 和 Wang (in press)。

4 PISA 2015 科学测评数据分析

4.1 研究问题与目的

通过对 PISA 2015 科学测评数据的分析，呈现出 MO-DINA 模型的现实需求和可应用性。基于上文中对科学素养所包含的三阶潜在结构划分，在针对科学素养的测评中，我们想测评被试在所有第一阶、第二阶和第三阶潜在特质(属性)上的表现情况。因此，本研究欲回答两个问题：(1) MO-DINA 模型是否适用于测评含三阶潜在结构的科学素养？如果可以，那么(2)科学素养的子维度中哪个对它的影响最大？即在 PISA 2015 中，科学素养的核心维度是哪个？

4.2 数据描述

4.2.1 多阶潜在特质设定

根据本文第 2 节的内容，PISA 2015 科学素养包含了三阶潜在结构，各阶潜在特质的名称及它们之间的结构关系见图 2。在数据分析时，我们依据 MO-DINA 模型将模型参数与多阶潜在特质进行匹配，第三阶潜在特质：$\theta^{(3)}$→科学素养；第二阶潜在特质：$\theta_1^{(2)}$→科学能力，$\theta_2^{(2)}$→科学知识，$\theta_3^{(2)}$→科学背景；第一阶潜在属性：A1→科学地解释现象，A2→评估和设计科学探究，A3→科学地解释数据和证据，A4→内容性知识，A5→过程性知识，A6→认知性知识，A7→个人背景，A8→地区/国家背景，A9→全球背景。需要说明的是，在第二阶潜在特质中，因为科学态度是通过学生问卷来获取的，并不包含在认知题目数据中，所以本研究暂不涉及。

4.2.2 被试与题目

根据《PISA 2015 技术报告》(OECD, 2017)的“附录 A：题池的分类(Item Pool Classification)”，数据清理过程如下：(1)选用“2015 field trial and main survey cluster”中 S01 所包含的 18 道题目，共 47548 人；(2)选用中国(QCH)样本，共 1079 人；(3)将数据中“not reached”和“no response”等设定为缺失值 NA；(4)删除在 18 题中全部缺失作答的 3 名被试，剩余 1076 人；(5)将剩余所有缺失值视为完全随机缺失。全贝叶斯 MCMC 算法可以根据其他参数的估计值计算出缺失值的后验分布，这是一种“自动填补”的过程，无需做其他设定。另外，DS519Q01 原为三级评分题目(i.e., $Y_{ni} \in \{0, 1, 2\}$)，限于 MO-DINA 模型暂只能处理二级评分题目，我们将该题目分数二级化：0→0, 1→0, 2→1。最终，清理后的数据包含 N = 1076 人在 I = 18 题上的二级评分数据。属性与题目之间的对应关系(i.e., Q 矩阵)见表 1。

表 1 PISA 2015 科学测验部分题目的 Q 矩阵

题目	$\theta^{(3)}$								
	$\theta_1^{(2)}$			$\theta_2^{(2)}$			$\theta_3^{(2)}$		
	A1	A2	A3	A4	A5	A6	A7	A8	A9
DS269Q01	1			1					1
DS269Q03	1			1					1
CS269Q04	1			1					1
CS408Q01	1			1				1	
DS408Q03	1			1				1	
CS408Q04	1			1				1	
CS408Q05		1			1			1	
CS521Q02	1			1			1		
CS521Q06	1			1			1		
DS519Q01		1				1	1		
CS519Q02	1			1			1		
DS519Q03			1	1			1		
CS527Q01			1			1			1
CS527Q03	1			1					1
CS527Q04	1					1			1
CS466Q01		1			1			1	
CS466Q07		1				1		1	
CS466Q05			1		1			1	

注：空白为“0”；选用“2015 field trial and main survey cluster” = S01 的题目。

4.3 分析

本研究选用 MO-DINA、HO-DINA 和 DINA 模型分别对该数据进行分析并比较。在潜在结构模型方面：对 MO-DINA 而言，其多阶潜在结构依据图 2 中结构设定(不考虑科学态度)；对于 HO-DINA 模型而言，假设第一阶属性直接受科学素养的影响，忽略第二阶潜在特质，即约束 $\gamma_m^{(2)} = 1$；对于 DINA 模型而言，忽略所有多阶潜在结构，直接使用无结构潜在结构模型。

三模型均使用两条马尔可夫链(随机起点)，每条链包含 10, 000 次迭代，其中预热 5, 000 次迭代，稀疏值 1。最终剩余 10, 000 次迭代用于参数估计。使用潜在量尺缩减因子(PSRF) (Brooks & Gelman, 1998)进行参数估计收敛检验，本研究中所有参数的 PSRF 均小于 1.2，表示参数估计已收敛。

本研究使用 AIC、BIC 和 DIC 作为模型−数据相对拟合指标，指标值越小的模型表明该模型与数据的拟合相对更好。另外，本研究使用后验预测模

型检验(posterior predictive model checking, PPMC)来评估模拟–数据绝对拟合指标, 其中后验预测概率(*ppp*), 接近 0.5 则表明模型与数据拟合, 小于 0.05 或大于 0.95 则表示该模型不拟合该数据。

4.4　结果

表 2 呈现了 3 个模型的各项模型–数据拟合指标值。首先, 根据 *ppp* 值, 3 个模型均拟合该数据。其次, 4 个相对拟合指标都判断 DINA 模型的相对拟合最差, 说明针对该数据应考虑高阶潜在结构。然后, 在 4 个相对拟合指标中, –2LL 和 AIC 均判断 MO-DINA 模型的相对拟合更好, 而 BIC 和 DIC 则判断 HO-DINA 模型的相对拟合更好, 这是由 BIC 和 DIC 对模型复杂性的惩罚相对更高导致的。另外, 由于 HO-DINA 模型是 MO-DINA 模型的特例(i.e., 约束 $\gamma_m^{(2)} = 1$), 似然函数比检验(Δ–2LL = 13, $df = 3$, $p < 0.05$)认为两模型差异显著, 应选择 MO-DINA 模型。最后, 再结合本研究的研究目的和问题, 我们综合认为 MO-DINA 模型更适宜于本研究。下文将基于 MO-DINA 模型的分析结果进行解读。

表 2　PISA 2015 科学测验部分题目数据的模型–数据拟合指标值.

模型	–2LL	AIC	BIC	DIC	*ppp*
MO-DINA	19332	19389	19673	24775	0.738
HO-DINA	19345	19399	19668	24644	0.716
DINA	19415	19962	22687	24856	0.692

表 3　PISA 2015 科学测验部分题目的参数估计值.

题目	g_i	s_i	95% CI (g_i)	95% CI (s_i)	IDI_i
DS269Q01	0.325	0.119	[0.263, 0.386]	[0.082, 0.158]	0.556
DS269Q03	0.459	0.070	[0.397, 0.521]	[0.042, 0.102]	0.471
CS269Q04	0.237	0.351	[0.190, 0.289]	[0.304, 0.398]	0.412
CS408Q01	0.434	0.181	[0.373, 0.489]	[0.142, 0.222]	0.385
DS408Q03	0.033	0.810	[0.015, 0.058]	[0.776, 0.843]	0.157
CS408Q04	0.429	0.261	[0.374, 0.487]	[0.219, 0.300]	0.310
CS408Q05	0.295	0.213	[0.220, 0.357]	[0.160, 0.266]	0.492
CS521Q02	0.548	0.133	[0.494, 0.602]	[0.097, 0.170]	0.319
CS521Q06	0.849	0.008	[0.809, 0.883]	[0.002, 0.017]	0.143
DS519Q01	0.106	0.524	[0.047, 0.163]	[0.457, 0.582]	0.370
CS519Q02	0.281	0.304	[0.231, 0.332]	[0.256, 0.353]	0.415
DS519Q03	0.323	0.228	[0.212, 0.404]	[0.174, 0.282]	0.449
CS527Q01	0.033	0.788	[0.012, 0.055]	[0.742, 0.831]	0.179
CS527Q03	0.393	0.330	[0.343, 0.442]	[0.289, 0.371]	0.277
CS527Q04	0.281	0.373	[0.203, 0.343]	[0.316, 0.423]	0.346
CS466Q01	0.448	0.182	[0.378, 0.514]	[0.140, 0.226]	0.370
CS466Q07	0.649	0.050	[0.543, 0.726]	[0.026, 0.080]	0.301
CS466Q05	0.342	0.243	[0.284, 0.398]	[0.184, 0.300]	0.415

注: 95% CI = 95%贝叶斯可信区间; g_i = 猜测参数, s_i = 失误参数; IDI_i = 题目区分度.

表 4　PISA 2015 科学测验部分题目的题目均值向量和方差协方差矩阵估计值.

	参数	后验均值	95% CI	相关系数
Σ	σ_β^2	1.773	[0.873, 3.571]	1.000
	$\rho_{\beta\delta}\sigma_\beta\sigma_\delta$	–1.833	[–3.719, –0.856]	–0.890
	σ_δ^2	2.394	[1.145, 4.778]	1.000
μ	μ_β	–0.783	[–1.408, –0.154]	
	μ_δ	–1.212	[–1.924, –0.493]	

表 3 呈现了题目参数的估计值。整体看这 18 道题的质量一般, 有个别题目的猜测参数或失误参数达到了 0.8 左右。这点根据题目区分度($IDI_i = 1 - s_i - g_i$) (de la Torre, 2008)也能够发现, 部分题目的区分度已经低于 0.2。这其中可能原因是(1)测验 Q 矩阵不完备(Köhn & Chiu, 2017); (2)题目涉及了 Q 矩阵以外的其他属性。另外, 表 4 呈现了 logit 转换后的题目参数的均值向量和方差协方差矩阵, 可以看到两类题目参数之间呈高程度负相关, 这符合 Zhan 等人(2018)的观点。

就高阶潜在特质的估计值而言, 首先, 1 个第三阶潜在特质和 3 个第二阶潜在特质的估计值整体分布形态基本一致, 这是因为它们之间的相关性较高(3 个回归系数分别为: 0.847 (SE = 0.094)、0.973 (SE = 0.025)和 0.927 (SE = 0.057), 因此, 它们之间相关系数约为 0.8)。需要说明的是, 特质之间在统计上有高相关并不一定代表它们是同一个特质。比如, 尽管身高和体重之间呈高相关, 但两者绝非同一种特质。因此, 当特质之间存在高相关时, 能否用一个笼统的高阶特质来囊括它们是需要做进一步理论判定的。基于 PISA 2015 科学素养测评框架, 我们认为这 3 个第二阶潜在特质在定义和内涵上都是不一样的, 不应将它们视为同一特质。另外, 我们还使用 HO-DINA 模型和单维两参数 Logistic 模型 (Birnbaum, 1968) 分析了该批数据, 发现 MO-DINA 模型中的第三阶潜在特质估计值与 HO-DINA 模型的高阶潜在特质估计值的相关系数为 0.996, 且与单维两参数 Logistic 模型的潜在特质估计值的相关系数为 0.936, 表明三者对“科学素养”的估计值具有高相关性, 同时也表明 MO-DINA 模型可提供更多的分析结果信息。

图 5 呈现了高阶潜在结构参数的估计值, 包括第三阶与第二阶潜在特质之间的回归系数和第二阶潜在特质与属性之间的属性区分度参数。首先, 3 个回归系数均接近于 1, 说明 PISA 2015 科学素养测评框架中把科学能力、科学知识和科学背景作为

科学素养的主要组成部分的做法是合理的。其次，根据这 3 个回归系数的大小可知：对科学素养而言，科学知识的影响最大，科学背景的影响次之，科学能力的影响最小。然后，根据属性区分度的大小可发现，(1)科学地解释现象对科学能力的影响最大；(2)过程性知识对科学知识的影响最大；(3)地区/国家背景对科学背景的影响最大。

表 5 呈现了个别被试的诊断结果示例。使用 MO-DINA 模型进行分析时，除了能够得到 9 个属性的诊断分类结果外，还能够得到被试在多阶潜在特质上的估计值。以 2 号和 23 号被试为例，尽管两者在属性模式上完全一样，但他们在多阶潜在特质上的表现还是有所差异的，说明它们对属性的掌握概率存在差异。

总体而言，根据对 PISA 2015 科学测验数据的分析结果，可以说 MO-DINA 模型满足本文的分析需求，在匹配 PISA 2015 科学素养测评框架的基础上，实现了对科学素养的客观测评。

5 模拟研究：参数估计返真性探究

5.1 研究设计与分析

在探讨完 MO-DINA 模型的现实可应用性后，我们通过一个简单的模拟研究来探讨它的参数估计返真性。模拟研究中的部分设定参考上文的实证数据分析结果，使用图 7 中的三阶潜在结构，即第三阶潜在特质 1 个，第二阶潜在特质 3 个，属性 $K = 9$ 个；题目数量设定为 $I = 30$, Q 矩阵设定见图 6；题目参数按如下方法生成：$(\text{logit}(g_i), \text{logit}(s_i))' = (\beta_i,$

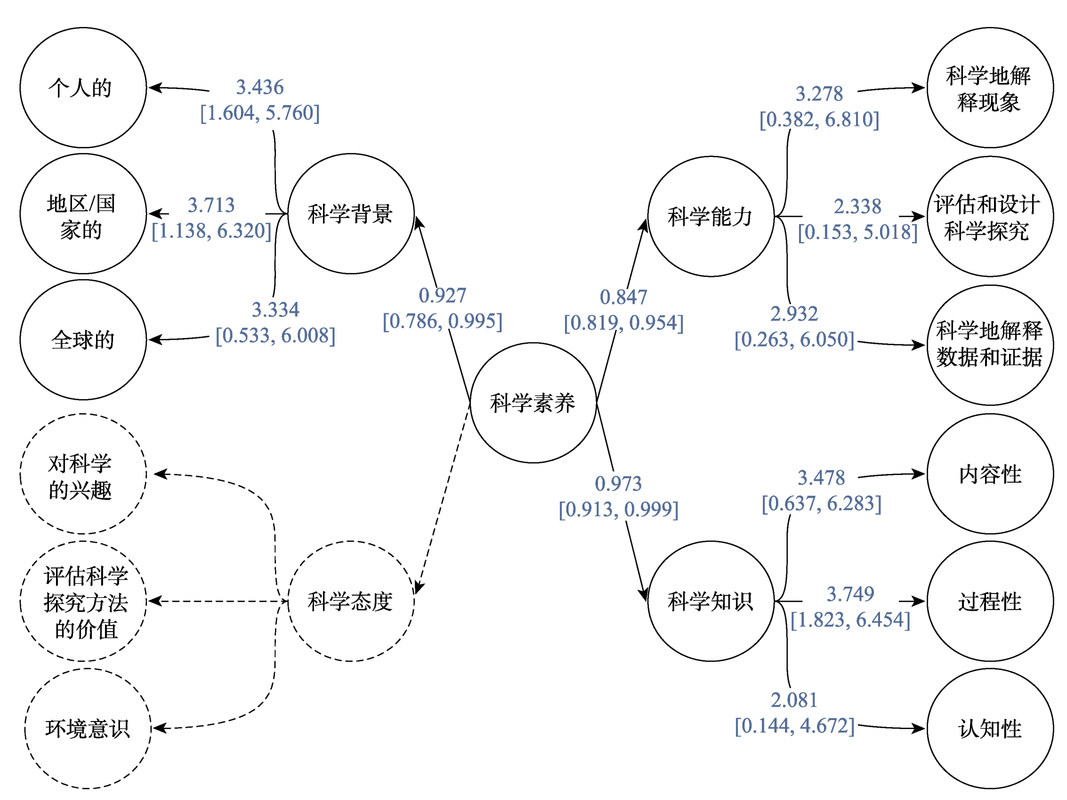


图 5 PISA 2015 科学测验中潜在结构参数估计值(基于 MO-DINA 模型).
注：括号内为 95%贝叶斯可信区间.

表 5 PISA 2015 科学测验部分题目数据的诊断结果示例(基于 MO-DINA 模型).

被试	α	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$	$\theta^{(3)}$
2	111111111	0.582 [–0.863, 2.194]	0.661 [–0.586, 2.174]	0.656 [–0.572, 2.175]	0.664 [–0.581, 2.194]
5	010001000	–0.873 [–2.317, 0.537]	–0.940 [–2.290, 0.276]	–0.910 [–2.307, 0.357]	–0.939 [–2.302, 0.263]
7	010000000	–0.919 [–2.429, 0.541]	–1.022 [–2.432, 0.198]	–1.028 [–2.445, 0.211]	–1.027 [–2.453, 0.183]
23	111111111	0.202 [–1.182, 1.950]	0.283 [–1.057, 1.961]	0.338 [–0.999, 1.959]	0.294 [–1.035, 1.968]
54	010101000	–0.831 [–2.414, 0.620]	–0.880 [–2.319, 0.461]	–0.870 [–2.368, 0.525]	–0.886 [–2.341, 0.426]
86	111101110	–0.404 [–2.082, 1.368]	–0.462 [–2.054, 1.314]	–0.468 [–2.034, 1.293]	–0.467 [–2.062, 1.300]

注：括号内为 95%贝叶斯可信区间.

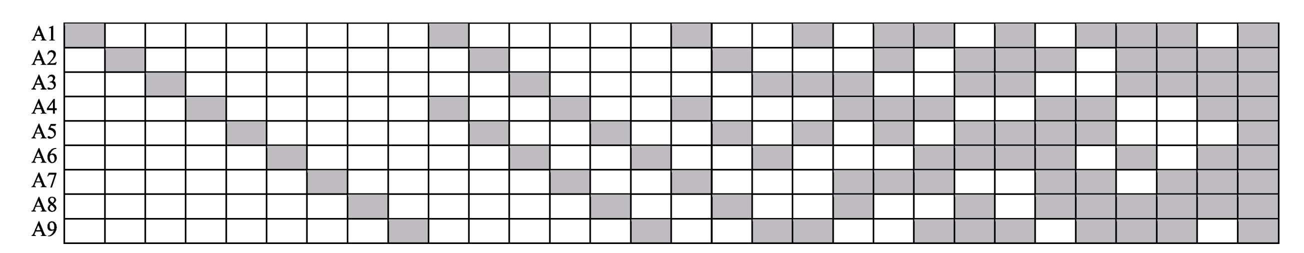


图 6　模拟研究中的 $K \times I$ 的 Q' 矩阵. 灰色表示“1”, 白色表示“0”.

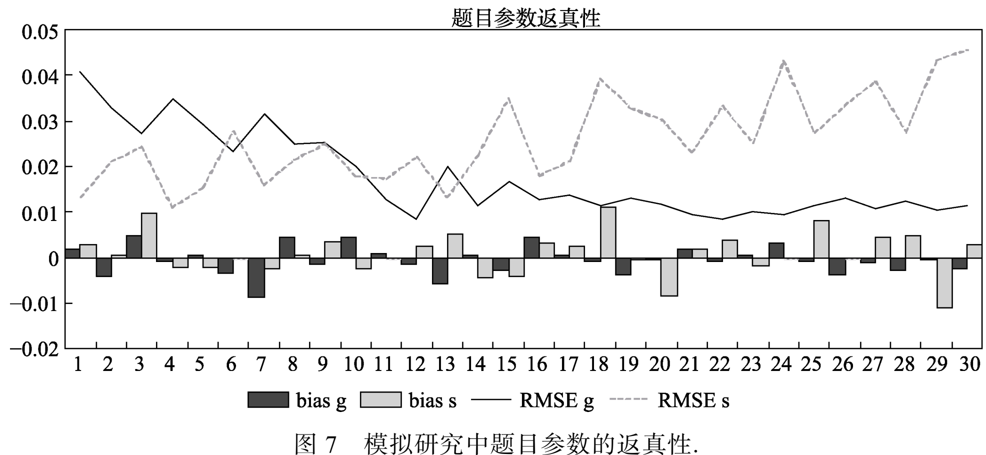


图 7　模拟研究中题目参数的返真性.
注: bias = 偏差; RMSE = 均方根误差.

$\delta_i)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 其中 $\mu_\beta = \mu_\delta = -2.197$, $\boldsymbol{\Sigma} = [1, -0.6; -0.6, 1]$; 属性截距向量 $\boldsymbol{\lambda}_0 = (-1, 0, 1, -1, 0, 1, -1, 0,1)$, 所有属性区分度均设定为 $\lambda_{1mk} = 1.5$, 即假设属性之间为中等程度相关; 被试量设定为 $N = 1,000$, 第三阶潜在特质从标准正态分布中生成, 第三阶与第二阶潜在特质之间的3个载荷均设定为 $\gamma_m^{(2)} = 0.8$, 即假设各二阶潜在特质之间相关系数为0.64。模拟研究中, 迭代次数、预热次数等与实证研究中的保持一致, 本研究中所有参数的 PSRF 均小于 1.2, 表示参数估计已收敛。另外, 使用偏差(Bias)、均方根误差(RMSE)和皮尔逊相关系数(Cor)来探究连续变量(e.g., 题目参数, 潜在特质)的返真性。使用属性正确判准率(ACCR)和属性模式正确判准率(PCCR)来探究属性的返真性。

5.2　结果

图 7 呈现了题目参数返真性。就 Bias 而言, 绝大多数题目的参数 Bias 小于 0.01, 猜测参数和失误参数的 Bias 的平均绝对值分是 0.002 和 0.004。就 RMSE 而言, 所有题目参数的 RMSE 均小于 0.05, 猜测参数和失误参数的 RMSE 的均值分别是 0.018 和 0.026。还可发现, 猜测参数的 RMSE 随着题目测查的属性数量的增加而下降, 而失误参数的 RMSE 随着题目测查的属性数量的增加而增加, 这与以往一些研究的结论是一致的(e.g., de la Torre, 2009; Zhan, Jiao, Liao, et al., 2018)。此外, 猜测参数和失误参数的 Cor 分别是 0.981 和 0.964, 即题目参数的估计值与真值之间呈高相关。整体而言, MO-DINA 模型的题目越参数返真性较好。

图 8 呈现了属性参数的 ACCR。9 个属性的 ACCR 均高于 0.900, 表明单个属性的参数估计返真性很好。另外, PCCR 为 0.512, 考虑到属性数量为 9, 即有 512 种可能的属性模式需要被估计, 根据已有研究经验, 该判准率符合预期。

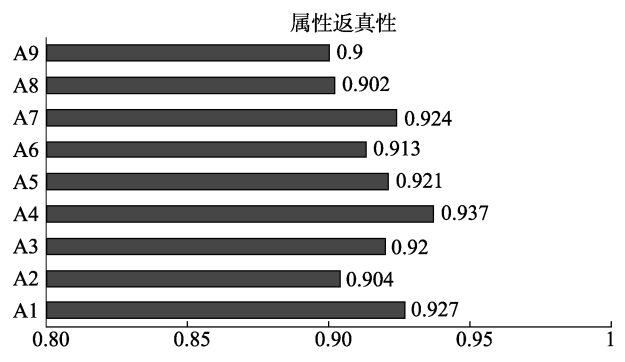


图 8　模拟研究中属性参数的属性正确判准率(ACCR).

表 6 呈现了高阶潜在特质参数的返真性。首先, 4 个高阶潜在特质的返真性类似, 1, 000 名被试的 bias 的平均绝对值约为 0.1, RMSE 的均值约为 0.69, Cor 均高于 0.7。参考以往关于 HO-DINA 模型的研究结果(e.g., de la Torre & Douglas, 2004; de la Torre, 2009; Zhan et al., 2018), 整体而言, 高阶潜在特质参数的返真性良好, 满足实际应用需求。

表 6 模拟研究中高阶潜在特质参数的返真性.

参数	tbias				RMSE				Cor
	平均绝对值	标准差	最小值	最大值	平均值	标准差	最小值	最大值	
$\theta^{(3)}$	0.100	0.124	–0.380	0.368	0.686	0.090	0.408	0.983	0.721
$\theta_1^{(2)}$	0.100	0.125	–0.378	0.352	0.689	0.092	0.385	0.983	0.719
$\theta_2^{(2)}$	0.104	0.126	–0.372	0.351	0.683	0.089	0.416	0.947	0.726
$\theta_3^{(2)}$	0.104	0.130	–0.481	0.381	0.690	0.095	0.358	1.050	0.715

注：bias = 偏差; RMSE = 均方根误差; Cor = 皮尔逊相关系数.

表 7 模拟研究中潜在结构参数的返真性

参数	bias				RMSE				Cor
	平均绝对值	标准差	最小值	最大值	平均值	标准差	最小值	最大值	
λ_{0k}	0.042	0.048	–0.066	0.072	0.189	0.062	0.129	0.305	0.982
λ_{1km}	0.116	0.051	0.015	0.172	0.346	0.057	0.245	0.429	0.982
$\gamma_1^{(2)}$	–0.031				0.053				
$\gamma_2^{(2)}$	–0.012				0.076				
$\gamma_3^{(2)}$	–0.012				0.076				

注：bias = 偏差; RMSE = 均方根误差; Cor = 皮尔逊相关系数; λ_{0k} = 属性难度参数, λ_{1km} = 属性区分度参数, γ_1, γ_2, γ_3 = 第三阶与第二阶潜在特质之间的回归系数.

表 7 呈现了高阶潜在结构参数的返真性。首先，对于属性难度参数的返真性优于属性区分度参数的返真性，与以往关于 HO-DINA 模型的研究结论一致。其次，第三阶潜在特质与 3 个第二阶潜在特质之间回归系数的返真性也较好，RMSE 均小于 0.08。整体而言，潜在结构参数的返真性较好。

6 总结与讨论

为实现对科学素养的客观且准确的测评，本文首先根据 PISA 2015 科学素养测评框架，提出了科学素养所包含的三阶潜在结构。然后，鉴于当前尚未有 CDM 能够处理包含三阶潜在结构的数据，我们提出了多阶认知诊断建模思路，并以 DINA 模型为例，建构了多阶 DINA (MO-DINA)模型。新模型采用全贝叶斯 MCMC 算法实现参数估计。新模型与 PISA 2015 科学素养测评框架相匹配，满足对科学素养的客观且准确测评的需求。之后，本文以 PISA 2015 科学测验数据分析为例来说明新模型的现实需求和可应用性。最后，通过一个模拟研究来探究新模型的参数估计返真性。实证研究结果表明当测验数据结构存在多阶潜在结构或者数据分析者需要了解被试在多阶潜在特质方面的表现时，可考虑使用 MO-DINA 模型。模拟研究结果表明本文提出的全贝叶斯 MCMC 算法能够为 MO-DINA 模型提供较好的参数估计返真性。

本文中，尽管 MO-DINA 模型是针对 PISA 2015 科学素养所包含的三阶潜在结构而提出的，且因为 MO-DINA 模型是 HO-DINA 模型的拓广，所以理论上该模型也可以适用于其他包含二阶及以上阶潜在结构的测验，比如国际数学和科学趋势研究(TIMSS)和(中国)国家义务教育阶段教育质量监测等大规模测验均包含了多阶潜在结构。当然，本研究并不是为了说明任何包含多阶潜在结构的测验或者任何针对科学素养的测验都需要使用 MO-DINA 模型来进行分析，而只是从“为学习而评价(assessment for learning)”的新测评理念出发，向读者提供一种新的测评视角和方法，以期进一步丰富数据分析模型的可选项。在实践中，我们除了可根据测验编制的理论和实际测验需求等来选择分析模型外，还可以尝试使用数据驱动方法，依据模型−数据拟合指标(e.g., AIC、BIC 和 DIC 等)来选择合适的模型，进而得到客观的、准确的以及满足需要的数据分析结果。

需要强调的是，一般存在 3 个及以上的低阶潜在特质时才会考虑使用高阶模型。具体而言，对于二阶 LSM (见式(1))，当 $K = 3$ 时，使用无结构潜在结构模型需要估计 $2^3 - 1 = 7$ 个结构参数，而使用二阶 LSM 仅需要估计 6 个参数(包含 3 个属性区分度和 3 个属性难度)；而对于第三阶与第二阶潜在特质而言，当第二阶潜在特质属性数量为 3 时，直

接估计 3 者之间的相关系数和估计第三阶与第二阶潜在特质之间的载荷均需要 3 参数，而当第二阶潜在特质数量大于 3 时，则使用高阶结构可以减少待估计参数数量。比如，就图 5 的三阶潜在结构而言，直接使用 DINA 模型需要估计 $2^9 - 1 = 511$ 个结构参数，使用 MO-DINA 模型仅需要估计 21 个结构参数(包含 9 个属性区分度、9 个属性难度和 3 个载荷)，可以大幅降低待估计参数数量。但若使用包含三个维度的二阶 DINA 模型，则同样需要估计 21 个结构参数(包含 9 个属性区分度、9 个属性难度和 3 个相关系数)，但此时就无法实现对“科学素养”维度的测量。因此，是否选用高阶模型，可以从理论(测验框架)和模型简约两个角度进行考虑，但究竟高阶模型是否合理，最终还要回归到理论，因为并不是所有潜在特质都适合建构高阶结构。比如，大五人格的五个维度就不应用高阶潜在特质“性格”去解释，因为从理论上讲人格的五个维度应该是独立的(尽管数据分析结果会存在低相关)。

尽管本研究将科学素养划分为了三阶潜在结构，但第一阶的属性粒度仍然较大，而通常 CDA 可能更适用于测评一些粒度较小的属性(see Leighton & Gierl, 2007; 詹沛达等, 2016)。实际上，基于 PISA 2015 科学测评框架，本研究中的第一阶属性还能够进一步划分为粒度更小的概念，比如，A1“科学地解释现象”还能够进一步划分为“回忆并应用适当的科学知识(Recall and apply appropriate scientifc knowledge)”和“提供解释性假设(Offer explanatory hypotheses)”等小粒度概念，详见 OECD (2016)的表 2.4a。尽管理论上我们可以使用包含四阶潜在结构的 MHO-DINA 模型做进一步分析，但受限于《PISA 2015 技术报告》中并未呈现题目与小粒度概念之间的具体对应关系(即没有相应的 Q 矩阵)，所以本文暂只关注到对科学素养所包含的三阶潜在结构的测评。另外，如有需要，后续还可以尝试使用三阶 IRT 模型(e.g., Huang et al., 2013)来分析该数据，并与本文的分析结果进行对比研究。

当然，由于能力和精力有限，本研究仍有一些局限值得后续做出进一步探究，比如：(1)尽管本文主要关注的是潜在结构模型，但仍仅使用了 DINA 模型作为测量模型，后续可尝试探究基于其他测量模型时的性能；(2)未考虑属性之间可能存在的层级结构(Leighton, Gierl, & Hunka, 2004)，如何将属性层级结构引入到多阶潜在结构中值得今后进一步关注(e.g., Zhan, Ma, Jiao & Ding, in press)；(3)仅涉及二分属性，而未考虑更为精细的多分属性(Karelitz, 2004)，如何将 MO-LSM 拓广到多分属性是一个有意义的话题(e.g., Zhan, Wang et al., in press)；(4)假设多阶潜在结构建构合理，而现实测验中多阶潜在结构的界定可能会存在偏差，在这种情况下 MO-DINA 模型的表现情况值得做进一步研究；(5) MO-DINA 模型仅考虑了单一的作答数据源，并未考虑诸如题目作答时间、鼠标点击次序数据等过程性数据，如何将过程性数据引入到当前建模思路中非常值得关注(e.g., Liu, Liu, & Li, 2018; Zhan et al., 2018)；(6) MO-DINA 模型仅针对横断测验数据，暂无法处理纵向测验数据，后续可尝试对其做进一步拓广(e.g., Li, Cohen, Bottge, & Templin, 2016; Zhan, Jiao, Liao & Li, in press)；(7)实证数据分析中，未考虑科学态度维度，如何将由学生问卷测评的科学态度和由认知题目测评的其他 3 个维度一同纳入到对科学素养的测评中值得今后做进一步探索。

参　考　文　献

Birnbaum, A. (1968). *Some latent trait models and their use in inferring a student's ability*. In F. M. Lord & M. R. Novick (Eds.). Statistical theories of mental test scores. Addison-Wesley, Reading, MA.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*. 7(4), 434–455.

Chen, F., Zhan, P., Wang, L., Chen, C., & Cai, M. (2015). The development and application of higher-order item response models. *Advances in Psychological Science, 23*, 150–157.

[陈飞鹏, 詹沛达, 王立君, 陈春晓, 蔡毛. (2015). 高阶项目反应模型的发展与应用. *心理科学进展, 23*, 150–157.]

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115 –130.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Song, H. (2009). Simultaneously estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620–639.

Gao, H. B. (2011). Results of the eighth survey on Chinese citizens' scientific literacy were released. *Bulletin of National Natural Science Foundation of China, 25*, 63–64.

[高宏斌. (2011). 第八次中国公民科学素养调查结果发布. *中国科学基金, 25*, 63–64.]

Hu, Y., Yang, S., & Lu, K. (2012). The research of assessment tools of adolescents' scientific literacy and its quality analysis. *Education Research Monthly, 3,* 16–21.

[胡咏梅, 杨素红, 卢珂. (2012). 青少年科学素养测评工具研发及质量分析. *教育学术月刊, 3,* 16–21.]

Huang, H.-Y., Wang, W.-C., Chen, P.-H., & Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*(8), 619–637.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign

Kang, C., & Xin, T. (2010). New development in test theory: multidimensional item response theory. *Advances in Psychological Science, 18*(3), 530–536

[康春花, 辛涛. (2010). 测验理论的新发展：多维项目反应理论. *心理科学进展, 18*(3), 530–536.]

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*(1), 112–132

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(1), 205–237.

Li, F., Cohen, A., Bottge, B, & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*(2), 181–204.

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology, 9, 1372.*

Liu, K., Li, C. (2015). The content and characteristic of PISA 2015 draft science framework. *Comparative Education Review, 37*(7), 98–105.

[刘克文, 李川. (2015). PISA 2015 科学素养测试内容及特点. *比较教育研究, 37*(7), 98–105.]

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics, 2*(2), 99–120.

Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus*, *112*(2), 29–48.

OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*. Paris: PISA, OECD Publishing

OECD. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: PISA, OECD Publishing

OECD. (2017). *PISA 2015 Technical Report*. Paris: PISA, OECD Publishing

Qin, H., & Qian, Y. (2008). A survey report on Shanghai adolescents' scientific literacy. *Research in Educational Development,* (24), 31–35.

[秦浩正, 钱源伟. (2008). 上海青少年科学素养调查报告. *教育发展研究,* (24), 31–35.]

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Ren, L., Zhang, C., & He, W. (2013). Constructing and anallysis of the model of how the factors affect the scientific literacy of Chinese citizens and a comparative investigation. *Studies in Science of Science, 31*, 983–990.

[任磊, 张超, 何薇. (2013). 中国公民科学素养及其影响因素模型的建构与分析. *科学学研究, 31*(7), 983–990.]

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 39*(4), 235–256.

Roos, J. M. (2014). Measuring science or religion? A measurement analysis of the National Science Foundation sponsored science literacy scale 2006–2010. *Public Understanding of Science, 23*(7), 797–813.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). *Assessment, 10*(2), 151–159.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*(7), 559–574.

The Ministry of Education of the People's Republic of China. (2017). *Compulsory education primary school curriculum standards.* Retrieved June 2, 2017, from http://www.moe.edu.cn/srcsite/A26/s8001/201702/t20170215_296305.html

[中华人民共和国教育部. (2017). *义务教育小学科学课程标准*. 2017-06-02 取自 http://www.moe.edu.cn/srcsite/A26/s8001/201702/t20170215_296305.html]

Tu, D., Cai, Y., & Ding, S. (2012). *Cognitive diagnosis: Theory, Methods, and Applications*. Beijing: Beijing Normal University Publishing Group.

[涂冬波, 蔡艳, 丁树良. (2012). *认知诊断理论、方法与应用.* 北京：北京师范大学出版社。]

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*(5), 295–316.

Zhan, P., Chen, P., & Bian, Y. (2016). Using confirmatory compensatory multidimensional IRT models to do cognitive diagnosis. *Acta Psychologica Sinica, 48*(10), 1347–1356.

[詹沛达, 陈平, 边玉芳. (2016). 使用验证性补偿多维 IRT 模型进行认知诊断评估. *心理学报, 48*(10), 1347–1356.]

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262–286.

Zhan, P., Jiao, H., Liao, D., & Li, F. (in press). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics.*

Zhan, P., Jiao, H., Liao, M., & Bian, Y. (2018). Bayesian DINA modeling incorporating within-item characteristic dependency. *Applied Psychological Measurement*. Advanced online publication. URL https://doi.org/10.1177/0146621618781594

Zhan, P., Jiao, H., Man, K., & Wang, L. (in press). Using

JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*.

Zhan, P., Ma, W., Jiao, H., & Ding, S. (in press). A sequential higher-order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Applied Psychological Measurement*.

Zhan, P., Wang, W.-C., & Li, X. (in press). A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. *Journal of Classification*.

Zhan, P., Wang, W.-C., & Wang, L. (2013). Testlet response theory: an introduction and new developments. *Advances in Psychological Science, 21*(12), 2265–2280.

[詹沛达，王文中，王立君. (2013). 项目反应理论新进展之题组反应理论. *心理科学进展, 21*(12), 2265−2280.]

附录：

1. MO-DINA 模型

测量模型选用引入题目内特征依赖性的贝叶斯 DINA 模型(Zhan, Jiao, Liao, & Bian, 2018)，可表示为：

$$P(Y_{ni}=1\,|\,\boldsymbol{\alpha}_n,\boldsymbol{\Psi}_i)=\frac{\exp(\beta_i)}{1+\exp(\beta_i)}+\left(1-\frac{\exp(\delta_i)}{1+\exp(\delta_i)}-\frac{\exp(\beta_i)}{1+\exp(\beta_i)}\right)\prod_{k=1}^{K}\alpha_{nk}^{q_{ik}},$$

$$\boldsymbol{\Psi}_i=\begin{pmatrix}\beta_i\\ \delta_i\end{pmatrix}\sim MVN_2(\boldsymbol{\mu},\ \boldsymbol{\Sigma}),$$

式中，Y_{ni}为被试 n 作答题目 i 的结果；$\boldsymbol{\Psi}_i=(\beta_i, \delta_i)'$为 logit 量尺上满足二元正态分布的题目参数向量(两者通常为负相关)，它们与常规 DINA 模型中的猜测和失误参数之间的关系为：$\mathrm{logit}(g_i)=\beta_i$, $\mathrm{logit}(s_i)=\delta_i$; q_{ik}为 Q 矩阵中元素，$q_{ik}=1$ 表示题目 i 考查了属性 k，反之，$q_{ik}=0$。将该模型与正文中式(4)相结合即可得到 MO-DINA 模型。

2. MO-DINA 模型中各待估计参数的先验分布设定如下：

首先，基于局部独立性假设，$Y_{ni}\sim$ Bernoulli $(P(Y_{ni}=1\,|\,\boldsymbol{\alpha}_n, \boldsymbol{\Psi}_i))$, $\alpha_{nk}\sim$ Bernoulli$(P(\alpha_{nk}=1\,|\,\boldsymbol{\theta}_n^{(2)}))$。

其次，关于题目参数的先验分布，参考 Zhan, Jiao, Liao 等人(2018)，设定如下：

$$\begin{pmatrix}\mathrm{logit}(g_i)\\ \mathrm{logit}(s_i)\end{pmatrix}=\begin{pmatrix}\beta_i\\ \delta_i\end{pmatrix}\sim N(\boldsymbol{\mu},\ \boldsymbol{\Sigma}),$$

$\boldsymbol{\mu}=(\mu_\beta, \mu_\delta)'$为 logit 转换后的题目参数均值，$\boldsymbol{\Sigma}$ 为方差协方差矩阵，有

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_\beta^2 & \\ \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \sigma_\delta^2\end{pmatrix},$$

$\rho_{\beta\delta}$为 logit 转换后的题目参数之间的相关系数。其中，μ_β 和 μ_δ 的超先验(hyper-prior)分布分别设定为 $\mu_\beta\sim N(-1.096, 4)$ 和 $\mu_\delta\sim N(-1.096, 4)$，鉴于 $\mathrm{logit}(-1.096)\approx 0.25$，所以该设定与四则一选择题的理论猜测概率相符合；另外，设定 $\boldsymbol{\Sigma}\sim$ InvWishart$(\mathbf{R}, 2)$，其中 $\mathbf{R}$ 为二维单位矩阵。

再有，关于高阶潜在特质参数，参考 Huang 等人(2013)，设定如下：

$$\theta_n^{(3)}\sim N(0,\ 1), \varepsilon_{nm}^{(2)}\sim N(0,\ 1-\gamma_m^{(2)}), \gamma_m^{(2)}\sim N(0.5,\ 0.25)\ I(-1,\ 1).$$

最后，关于高阶潜在结构参数，参考 Zhan, Jiao 和 Liao (2018)，设定如下：

$$\lambda_{0k}\sim N(0,\ 4), \lambda_{1mk}\sim N(0,\ 4)\ I(\lambda_{1mk}>0)$$

Using a multi-order cognitive diagnosis model to assess scientific literacy

ZHAN Peida; YU Zhaohui; LI Feiming; WANG Lijun

(*College of Teacher Education, Zhejiang Normal University, Jinhua* 321004, *China*)

Abstract

In PISA 2015, scientific literacy is defined as “the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen”. There are four interdependent dimensions are specified in the

scientific literacy assessment framework for PISA 2015: Competencies, Knowledge, Contexts, and Attitudes. Given that knowledge of scientific literacy contributes significantly to individuals' personal, social, and professional lives, it is of vital importance to find an objectively and accurately assessment method for scientific literacy. However, only unidimensional IRT models were used in the analysis in PISA 2015. Which means that the analysis model does not match with such a multidimensional assessment framework. It is desired to develop a new analysis model. This study attempts to measure scientific literacy in cognitive diagnostic assessment for the first time.

According to the scientific literacy assessment framework for PISA 2015, a third-order latent structure for scientific literacy is first pointed out. Specifically, the scientific literacy is treated as the third-order latent trait; Competencies, Knowledge, Contexts, and Attitudes are all treated as second-order latent traits; And nine subdomains, e.g., explain phenomena scientifically and content knowledge, were treated as first-order traits (or attributes). Unfortunately, however, there is still a lack of cognitive diagnosis models that can deal with such a third-order latent structure. To this end, a multi-order DINA (MO-DINA) model was developed in this study. The new model is an extension of the higher-order (HO-DINA) model, which is similar to the third-order IRT models. To illustrate the application and advantages of the MO-DINA model, a sub-data of PISA 2015 science assessment data were analyzed. Items were chosen from the S01 cluster, and participants were chosen from China. After data cleaning, 1076 participants with 18 items were retained. Three models were fitted to this sub-data and compared, the MO-DINA model, in which the third-order latent structure of scientific literacy was considered; the HO-DINA model, in which the scientific literacy was treated as a second-order latent trait and contacted with attributes directly; and the DINA model.

All three models appear to provide a reasonably good fit to data according to the posterior predictive model checking. According to the –2LL, AIC, BIC, and DIC, the DINA model fits the data worst, and the MO-DINA model fits the data best, the results of MO-DINA model are used to make further interpretations. The results indicated that (1) the quality of 18 items are not good enough; (2) The correlations among second-order latent traits are high (0.8, approximately); (3) Knowledge has the greatest influence on scientific literacy, Contexts second, and Competencies least; (4) Explain phenomena scientifically, procedural knowledge, and local/national has the greatest influence on Competencies, Knowledge, and Contexts, respectively. In addition, a simulation study was conducted to evaluate the psychometric properties of the proposed model. The results showed that the proposed Bayesian MCMC estimation algorithm can provide accurate model parameter estimation.

Overall, the proposed MO-DINA model works well in real data analysis and simulation study and meets the needs of assessment for PISA 2015 scientific literacy which included a third-order latent structure.

Key words scientific literacy; cognitive diagnosis; PISA; DINA model